Invited paper

Speech Segmentation Algorithm Based on an Analysis of the Normalized Power Spectral Density

Dzmitry Pekar and Siarhei Tsikhanenka

Belarusian State University, Minsk, Belarus

Abstract—This article demonstrates a new approach to speaker independent phoneme detection. The core of the algorithm is to measure the distance between normalized power spectral densities in adjacent, short-time segments and verify it based on velocity of changes of values of short-time signal energy analysis. The results of experiment analysis indicate that proposed algorithm allows revealing a phoneme structure of pronounced speech with high probability. The advantages of this algorithm are absence of any prior information on a signal or model of phonemes and speakers that allows the algorithm to be speaker independent and have a low computation complexity.

Keywords— phoneme segmentation, power spectral density, short-term signal energy, speaker independent, voice systems.

1. Introduction

Modern research in the field of speech technologies aimed at creating robust speech systems that can be used as a basis of voice interfaces for managing the information environment.

This article describes an algorithm for converting acoustic information in speech segments into phonemes. The algorithm is based on an analysis of speech spectral characteristics. Human speech is a sequence of phonemes that have their own unique spectral parameters: power spectral density (PSD) [1], the centers of gravity of the sub-bands of the spectrum [2], the slope spectrum [3], the fundamental frequency [4], formants [5], etc. An analysis of the normalized PSD of the temporal segment of the speech signal is used in the presented algorithm.

2. General Structure of the Algorithm

The general structural scheme of the proposed algorithm is presented in Fig. 1.

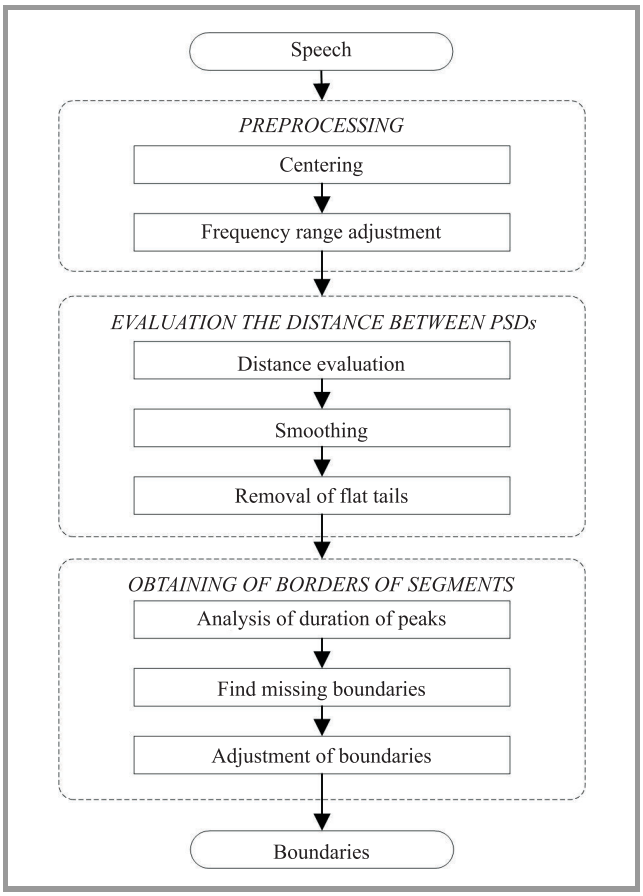

Fig. 1. General structural scheme of the algorithm.

3. Preprocessing

Preliminary signal processing is carried out to improve the characteristics of the signal, such as reducing inserted distortions and adjustment of the frequency range. The first step is centering the signal – removing the constant component introduced by the hardware of the system according to the expression:

$$\overline{s}_i = s_i - s_{mean}\,, \tag{1}$$

where

$$s_{mean} = \frac{\sum_{i=1}^{L} s_i}{L}\,, \tag{2}$$

where: $\overline{s}_i$ — samples of the corrected signal, s_i – samples of the original signal, L – number of samples.

The second step is pre-emphasizing which is performed to adjust frequency range and to strengthen the information contained in higher frequency components [6], [7]:

$$H(z) = 1 - 0.95z^{-1}. \tag{3}$$

4. Distance Evaluation between PSD Adjacent Short-term Segments

After completing speech signal preprocessing, the next step is carried out for the extraction of the numeric feature, which is the normalized power spectral density. Human speech has emotion, intonation, stresses, and therefore the signal takes significant fluctuations in energy, even within the same phoneme, which may lead to incorrect segmentation. Energy normalization is used to reduce the influence of the mentioned factors: the energy of each time segment should be equal to unity, which allows to take into account only the energy distribution in frequency, rather than its absolute value. Calculating the distance between PSDs is performed according to the algorithm, which is shown in Fig. 2.

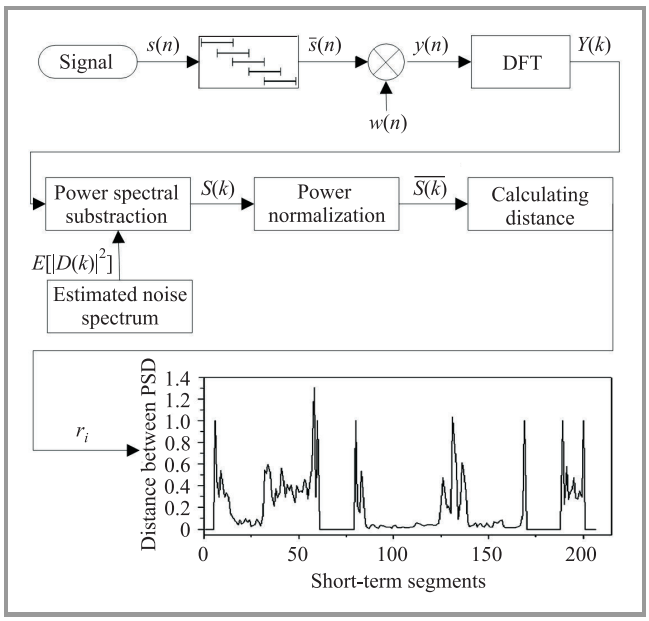

Fig. 2. Structural scheme of algorithm for computing the distance between PSDs.

In the first step the segmentation of the speech signal into portions with 50% overlap is carried out, which improves the temporal localization of spectral change. To reduce the influence of boundary effects arising from the segmentation of the signal, a weighted Chebyshev window is used, which is determined by the following expression:

$$w(n) = \sum_{k=0}^{n} \frac{(-1)^k (n-k-2)! z_0^{-2k}}{(N-n-k-1)!k!(n-k)!}, \tag{4a}$$

$$z_0 = \mathrm{ch}\left(\frac{1}{\mathrm{N}-1}\mathrm{arch}\frac{1}{\mathrm{h}}\right), \tag{4b}$$

where N – number of samples in short-term segment.

In Eqs. (4a) and (4b) the coefficients $a(n)$ are defined for n from 0 to L, where $L = N/2-1$ for an even N. The values are determined from the condition of even symmetry of the weight function. These coefficients are normalized by equating their sum to unity.

At the next step the time segment undergoes a discrete Fourier transform to obtain the signal spectrum:

$$Y(k) = \sum_{n=0}^{N-1} y(n)\exp(-j2\pi kn/N), \tag{5}$$

where: $y(n) = w(n)s(n)$ – weighted samples of speech signal, $k = 1,2,\ldots,N-1$ – spectral components number.

Taking into account the additive nature of noise, the signal in time domain can be represented as:

$$y(n) = s(n) + d(n), \tag{6}$$

where: $s(n)$ – source signal, $d(n)$ – noise.

Then, the power spectrum is described by the following expression [8]:

$$|Y(k)|^2 = |S(k)|^2 + |D(k)|^2 + S^*(k)D(k) + D^*(k)S(k), \tag{7}$$

where: $S(k)$ – spectrum of the source signal, $Y(k)$ – spectrum of the real measured signal, $D(k)$ – noise spectrum, which is calculated during pauses in the absence of speech signals, $S^*(k)$, $D^*(k)$ – imaginary part of the spectrum of not noisy signal and noise signal, respectively.

Terms on the right side of expression (7) can be estimated as $E[|D(k)|^2]$, $E[S^*(k)D(k)]$, $E[D^*(k)S(k)]$, where the complex assessment is eliminated assuming that the noise $d(n)$ is a zero mean and is not correlated with the signal $s(n)$.

Then the estimation is calculated according to the following expression:

$$|S(k)|^2 = |Y(k)|^2 - E[|D(k)|^2]. \tag{8}$$

In the next step the signal energy is normalized according to the expression:

$$\overline{S}(k) = \frac{S(k)}{\left(\sum_z |S(z)|^p\right)^{1/p}}, \tag{9}$$

where: $S(k)$ – original kth spectral component, $\overline{S_k^i}$ – normalized spectral component, p – parameter, which is $p = 2$.

Then the distance between ith and jth PSDs is defined by the expression:

$$r_i = \sum_k \left|\overline{S}(k)_i - \overline{S}(k)_j\right|^E, \tag{10}$$

where the exponent E is a parameter that determines the sensitivity of the algorithm to the spectral changes.

The final step aims to give a threshold to the distance between PSDs and to remove the small values from the characteristics:

$$\overline{r} = \begin{cases} 0, r < E[r] \\ r, r \geq E[r] \end{cases}, \tag{11}$$

where $E[r]$ – average distance between two PSDs, calculated across the signal.

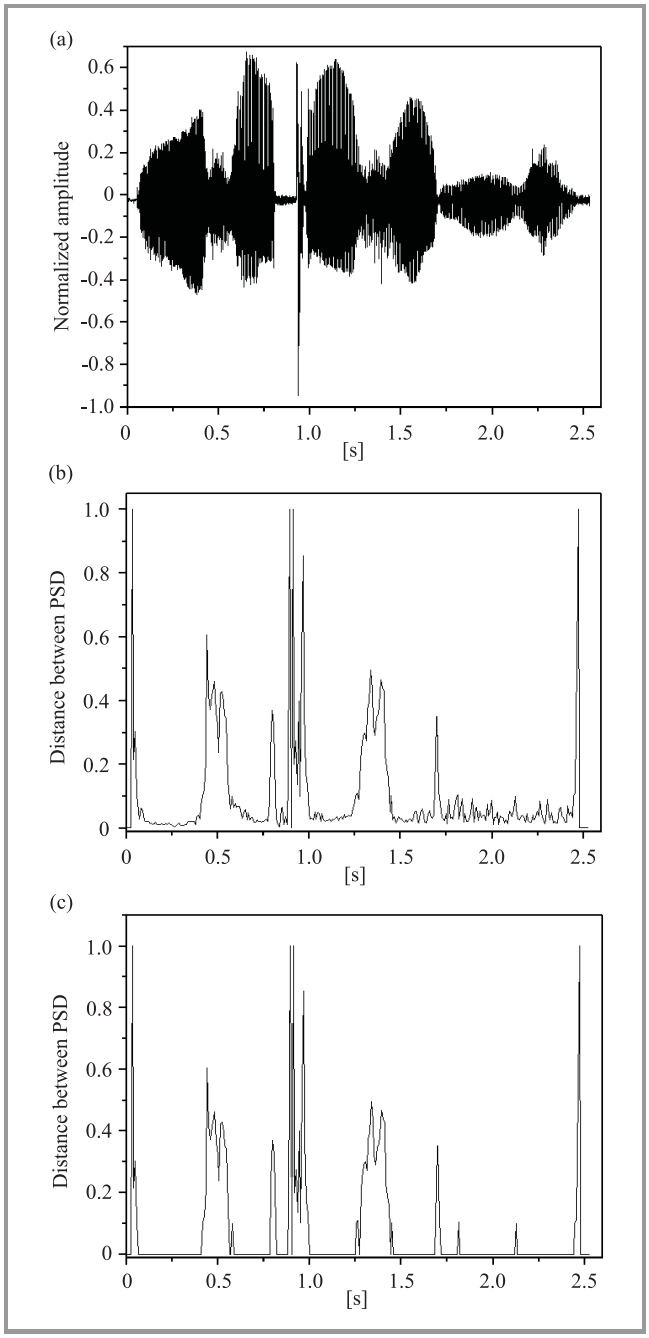

Fig. 3. The stages of calculating the distance between twp PSDs: (a) original signal, (b) distance between PSD, (c) distance between PSD after threshold.

Figure 3 shows the stages of calculating the distance between PSDs. There are three types of fragments in Fig. 3c: fragments of the long zero which correspond to quasi-stationary segments, broad emissions which are associated with noise segments and narrow peaks which are related to the transitions between the quasi-stationary segments or random spectral changes. Since the vocal tract has inertia, then the sequence of transition segments appears as closely spaced peaks which correspond to the restructuring of the vocal organs. To isolate them in a typical continued fragment, smoothing is performed by the method of least squares, which is similar to passing the signal through a low pass filter, which is determined by:

$$\widetilde{r}_k = \frac{1}{2L_f + 1} \sum_{n=L_f}^{L_f} r_{k-n} b_n \,, \tag{12}$$

where: L_f – length of the filter, in this particular case $L_f = 9$, b_n – filter coefficients. As the filter coefficients it is advisable to use the values: $b_n = \{(-21, 14, 39, 54, 59, 54, 39, 14, -21)/231\}$. This allows achieving the required smoothing while maintaining the necessary detail.

The application of smoothing shifts the edges of the peaks. To neutralize the mentioned effect the algorithm for analysis of the growth rate peak is used in accordance to which flats edges are cut which rate of change in each point Eq. (13) is less than average rate over all peak Eq. (14) multiplied by weighting factor *thr* (see Fig. 4):

$$r'_k = \frac{-\widetilde{r}_{k+2} + 4\widetilde{r}_{k+1} - 3\widetilde{r}_k}{2} \,, \tag{13}$$

$$E[r'] = \frac{1}{W-2} \sum_{i=2}^{W-1} r'_i \,, \tag{14}$$

where W – peak length.

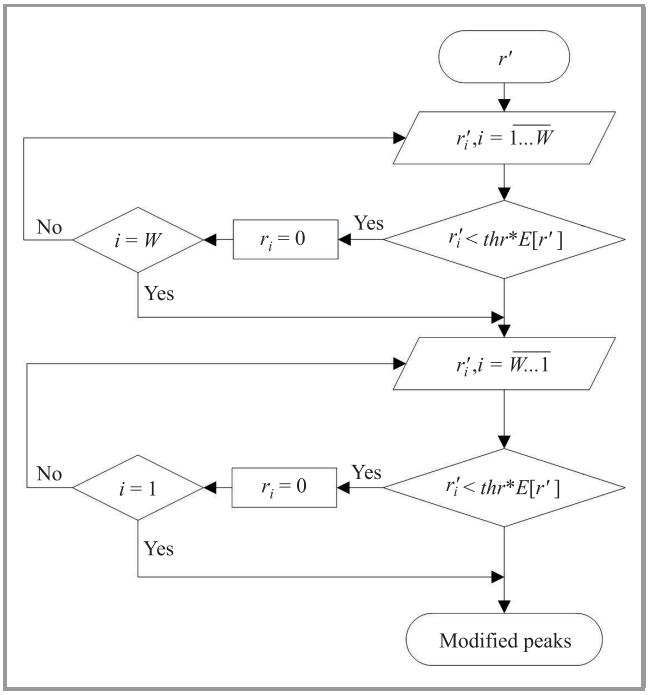

Fig. 4. Structural scheme of the analysis procedure of the edges of the peaks.

Next, the analysis of the edges of the peaks is performed, according to procedure in Fig. 4.

Experimental selected value of the weighting factor *thr* was in the range of 12%–17%. Figure 5 shows result of smoothing emissions.

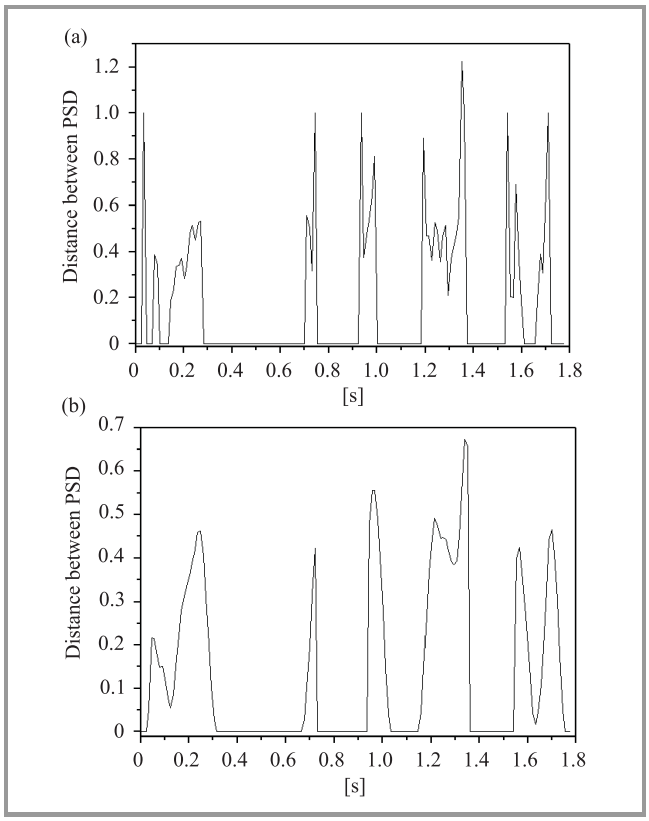

Fig. 5. The results of smoothing: (a) before smoothing, (b) after smoothing.

5. Delimitation Segments

The purpose of this stage is to define the representatives of smoothed PSD, representing individual phonemes and the transition boundaries between phonemes. In first step, the duration of the peaks is analyzed, according to the algorithm presented in Fig. 6.

In the above algorithm: L_{peak} – the duration of peak, $L_{\min}$ – the minimum allowable length of the segment. A phoneme length takes values from 50 ms to 250 ms [9], therefore, 50 ms can be taken as the minimum acceptable duration of the segment. When the duration exceeds the peak threshold, it exhibits two boundaries on the peak width, limiting the found segment, otherwise, put up one, to the maximum peak.

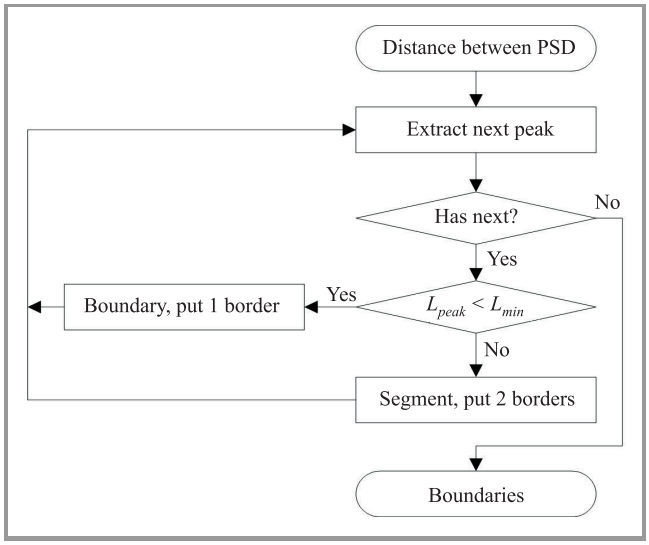

Fig. 6. Structural scheme of algorithm analysis of the duration of the peaks.

After a preliminary determination of boundaries a procedure that checks for the presence of missed boundaries is performed. Based on testing of different criteria for deciding on the loss of the border, the most efficient test was the difference between the lengths of the two largest segments. If the longest segment has a duration of twice or more greater than the second length of the segment, it is necessary to check first on the presence of a merger of phonemes. Since the analysis in the frequency domain has not given the significant differences in the characteristics, then further analysis is conducted in the time domain, where the analyzed characteristic is the rate of change of short-time signal energy, which is determined according to the algorithm shown in Fig. 7.

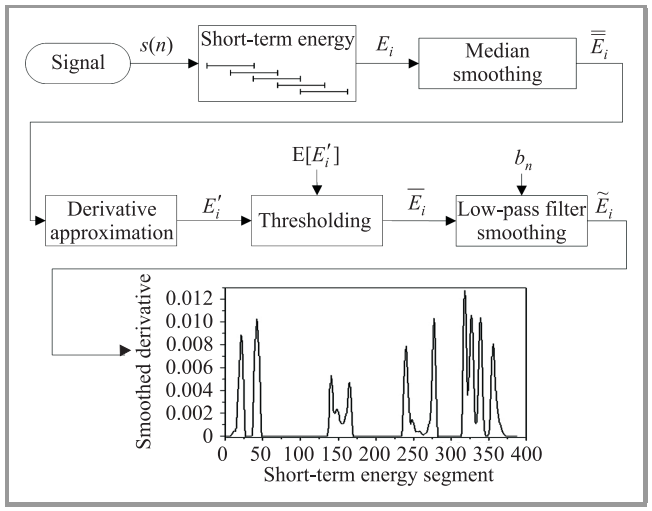

Fig. 7. Structural scheme of the algorithm for evaluation values of rates of change in short-term signal energy.

Calculation of short-time energy is carried out according to the expression:

$$E_i = \sum_{j=0}^{N/2} s_{i,j}^2 , \tag{15}$$

where: E_i – energy of the ith time segment, $s_{i,j}$ – jth count in the ith time segment.

Median filter removes random emissions, makes the curve of the values E_i smoothed and allows avoiding big leaps in rate of change E'_i of short-time energy E_i which is calculated similar to the expression (13). Threshold processing $\overline{E}_i$ is conducted according to expression (11) with the threshold $E[E']$ defined by the expression (14). To isolate specific fragments $\widetilde{E}_i$ curve of rate of change of signal energy, all values are passed through a low pass filter Eq. (12). After these steps, the output is a similar to that which arises in the analysis of the distance between PSDs, which allows the use of an algorithm to analyze the duration of the peaks

as shown in Fig. 6. The detected segment or border of an interval occupied with the checked fragment is added to set of borders and segments that have been found before.

6. Boundaries Adjustment

The repeated processing of the values of distances between PSDs leads to a shifting of maximum peaks and the need to adjust the positions of boundaries of detected segments. The speech signal does not undergo significant changes for up to 20–25 ms, therefore, its spectral composition does not undergo significant changes either.

Then, the estimated boundary can occupy the position as at the beginning of the quasi-stationary section, and in the end. Therefore it is necessary to adjust boundaries in the range $[t_c - 20 \text{ ms};\ t_c + 20 \text{ ms}]$, where t_c – the current position of the border segment. The position is accepted as more refined, if the following equation is fulfilled:

$$r_i^{new} > r_i^{old}, \tag{16}$$

where: r_i^{new} – distance between PSDs in the new position of the boundary, r_i^{old} – distance between PSDs in the old position of the boundary. Distances are calculated according to expression (12).

7. Experiments

The experimental validation of the algorithm was conducted to determine the probability of correct segmentation.

Table 1
The results of the segmentation algorithm

Property [%] / W_B [sample]	P_{right}^{8000}	P_{wrong}^{8000}	P_{right}^{11250}	P_{wrong}^{11250}	P_{right}^{16000}
$64_{8\text{ bit}}$	39,1	54,2	31,7	59,4	–
$64_{16\text{ bit}}$	41,5	48,2	36,3	57,2	–
$128_{8\text{ bit}}$	49,9	21,3	51,2	15,4	41,1
$128_{16\text{ bit}}$	75,7	11,8	56,4	17,8	49,6
$256_{8\text{ bit}}$	43,4	25,3	46,8	23,4	62,2
$256_{16\text{ bit}}$	50,1	18,4	63,3	19,5	**89,4**
$512_{8\text{ bit}}$	–	–	38,1	26,4	41,8
$5124_{16\text{ bit}}$	–	–	41,4	19,1	51,4
Property [%] / W_B [sample]	P_{right}^{8000}	P_{wrong}^{8000}	P_{right}^{11250}	P_{wrong}^{11250}	P_{right}^{16000}
$64_{8\text{ bit}}$	–	–	–	–	–
$64_{16\text{ bit}}$	–	–	–	–	–
$128_{8\text{ bit}}$	31,7	57,1	31,4	–	–
$128_{16\text{ bit}}$	27,4	63,1	28,7	–	–
$256_{8\text{ bit}}$	15,7	53,1	25,8	58,8	24,4
$256_{16\text{ bit}}$	**9,2**	68,4	21,1	67,1	19,9
$512_{8\text{ bit}}$	30,7	59,7	16,3	73,8	11,5
$5124_{16\text{ bit}}$	24,1	67,1	12,1	**90,7**	**8,9**

Denote the following notations:

- P_{right}^{F} – percentage of correctly exposed boundaries at the sampling rate equal to F [Hz]:

$$P_{right}^{F} = \frac{N_{right}}{N_{total}} 100\%, \tag{17}$$

where: N_{right} – number of correctly exposed borders, N_{total} – total count of boundaries which is calculated at the beginning of experiments.

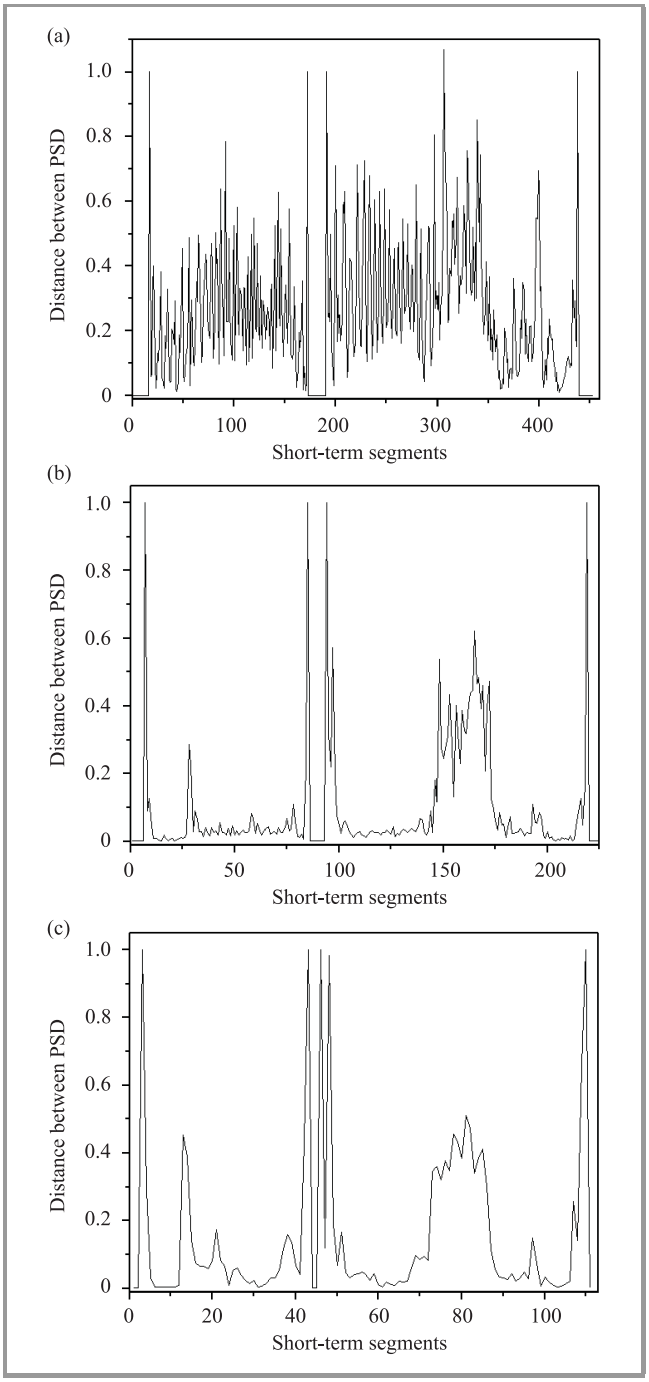

Fig. 8. The distance between PSDs depending on the length of the window at the sampling rate $F_s = 800$ Hz: (a) 64 samples, (b) 128 samples, (c) 256 samples.

- P^F_{wrong} – the relative number of wrongly exposed borders:

$$P^F_{wrong} = \frac{N_{wrong}}{N_{total}} 100\%, \tag{18}$$

where: N_{wrong} – number of wrongly exposed borders.

- P^F_{missed} – percentage of missed boundaries defined by the following expression:

$$P^F_{missed} = 1 - P^F_{right} \tag{19}$$

- W_B – length of analyzable short-term segment in samples with the B bits per sample ratio.

The experimental digital speech signals have a sampling frequency in the range from 8000 Hz to 32000 Hz. If it goes below this range, the completeness of the presentation of the speech signal is not achieved; and above this range, there is a redundancy of representation, and excessive increase in computational complexity.
The analysis of the results of the experiment (Table 1) shows that to achieve the best segmentation results it is necessary to choose a combination of settings. This fact is explained by the excessive reduction of the window length that leads to noise pollution characteristics of the distance between PSDs (Fig. 8a), which causes the appearance of false boundaries. That's why calculations of the values in the lower left side of the table were not carried out. Otherwise, a decreased temporal localization of spectral changes appears (Fig. 8b), which reduces the accuracy of segmentation.
In the course of the experiment 50 samples of the speech signal were tested which priori contained N_{total} boundaries. The samples were uttered by male and female speakers.

8. Conclusion

The proposed algorithm for segmentation of digital speech signals is based on the analysis of the normalized power spectral density which allows determination of the phonemic structure of pronounced speech with probability up to 90% without any priori information about signal, acoustic models of phonemes, or speaker individual characteristics which allows performing segmentation task of speech pronounced by different speakers.

References

[1] A. Saheli and A. Abolfazl, "Speech recognition from PSD using neural network", in *Proc. Int. MultiConf. Engin. Comp. Scient. IMECS 2009*, Hong Kong, 2009, vol. 1, pp. 174–176.

[2] B. Gajic and K. Paliwal, "Robust parameters for speech recognition based on subband spectral centroid histograms", in *Proc. 7th Eur. Conf. Speech Commun. Technol. EUROSPEECH 2001*, Aalborg, Denmark, 2001.

[3] C. Espy-Wilson and S. Manocha, "A new set of features for text-independent speaker identification", in *Proc. Int. Conf. Spoken Lang. Proces. INTERSPEECH 2006*, Pittsburgh, USA, 2006.

[4] P. Labutin and S Koval, "Speaker identification based on the statistical analysis of f0", in *Proc. 16th Annual Conference IAFPA 2007*, Plymouth, UK, 2007.

[5] T. Becker and M. Jessen, "Forensic speaker verification using formant features and gaussian mixture models", in *Proc. Int. Conf. Spoken Lang. Proces. INTERSPEECH 2008*, Brisbane, Australia, 2008.

[6] E. H. Kim and K. H. Hyun, "Robust emotion recognition feature, frequency range of meaningful signal", in *Proc. IEEE Int. Worksh. Robot Human Interact. Commun.*, Nashville, USA, 2005.

[7] M. A. Al-Alaoui and L. Al-Kanj, "Speech recognition using artificial neural networks and hidden Markov models", *IEEE Multidiscipl. Engin. Educ. Mag.*, vol. 3, no. 3, 2008.

[8] N. Bhatnagar, "A modified spectral subtraction method combined with perceptual weighting for speech enhancement", M.Sc. thesis, The University of Texas, Dallas, August 2002.

[9] M. S. Medvedev, "Ispolzovanije vejvlet-preobrazowanija dla postrojenia modelej fonem russkowo jazyka", *Wiestnik Sibirskogo Federalnogo Universiteta*, no. 9, p. 198, 2006 (in Russian).

Dzmitry Pekar was born in Svisloch, Belarus, in 1987. He received a specialist degree in radiophysics in 2010 from Department of Intelligent Systems and Networks, Belarusian State University, Belarus. Since then he has been working on a Ph.D. thesis concerning speech segmentation and speech recognition. His research interests includes GMM based and wavelet based Russian speech processing.
email: pekar.dima@gmail.com
Belarusian State University
Nezavisimosti av. 4
Minsk, Belarus, 220030

Siarhei Tsikhanenka was born in Gomel, Belarus, in 1983. He received B.Sc. degree, specialist degreed and Ph.D. degrees in 2004, 2005 and 2009 from Department of Intelligent Systems and Networks, Belarusian State University, Belarus. Since 2008 he has taken a lead scientist role at multimedia data processing laboratory at Department of Intelligent Systems and Networks, Belarusian State University, Minsk. His main scientific interests are still image compression, speech processing and enterprise resource planning systems implementation.
e-mail: siarhei.tsikhanenka@gmail.com
Belarusian State University
Nezavisimosti av. 4
Minsk, Belarus, 220030